A combined deep learning and structure based cheminformatic approach to understand ligand blockage activity on the hERG channel

Nathaniel Smith, David Bushiri Pwesombo, Han Sun

Leibnitz-Forschungsinstitut für Molekulare Pharmakologie, Robert-Rössle-Strasse 10, 13125 Berlin, Germany

Introduction

- Bioassays are used to determine ligand inhibition of a target gene with High-Throughput-Screening (HTS), often to find potential drugs.
- However setup and execution of assays is time consuming and resource intensive, depending on the type of screening. Primary screens may need 1000s of compounds just to find a hit.
- A series of supervised binary classification deep learning (DL) models were trained on a large dataset on the hERG (human *Ether-à-go-go* Related Gene). Due to the importance of preclinical testing of the hERG channel, large amounts of assay data is available for *in silico* techniques.
- These models were tested to predict ligand activity, and compared for accuracy and other metrics.

Aims

- To create a machine learning model that can accurately predict inhibitors of hERG.
- To compare the models to the results of molecular docking.

Docking Methods

- PubChem SDF structures downloaded by CID, hERG structure from RCSB (7CN1)[2].
- Ligands prepared using LigPrep, then Schrodinger HTVS, SP, XP (Glide).

ML Methods

Dataset

- 298,016 molecules were taken from the hERGCentral database and encoded via PaDEL molecular descriptors (2217 descriptors) or molecular graphs[1].
- The datasets were filtered to remove erroneous values, then split into train/test sets.
- Classification labels were set as 1, for inhibitor, 0 for non-inhibitor.

ML Configuration

- Random Forest (RF), Decision Tree (DT), 3-Layer Deep Neural Network (3L-DNN), and 3-Layer Graph Convolutional Network (3L-GCN) ML models were used.
- RF, DF models implemented using Scikit-Learn (v1.1.1) library.
- 3L-DNN implemented using TensorFlow (v2.9.1)
- GNNs implemented using PyTorch (v1.10.2), and PyTorch Geometric (v2.0.4).
- Accuracy and Confusion Matrix (CM) metrics used for binary classification, with the classification threshold set to 0.5.

Workflow

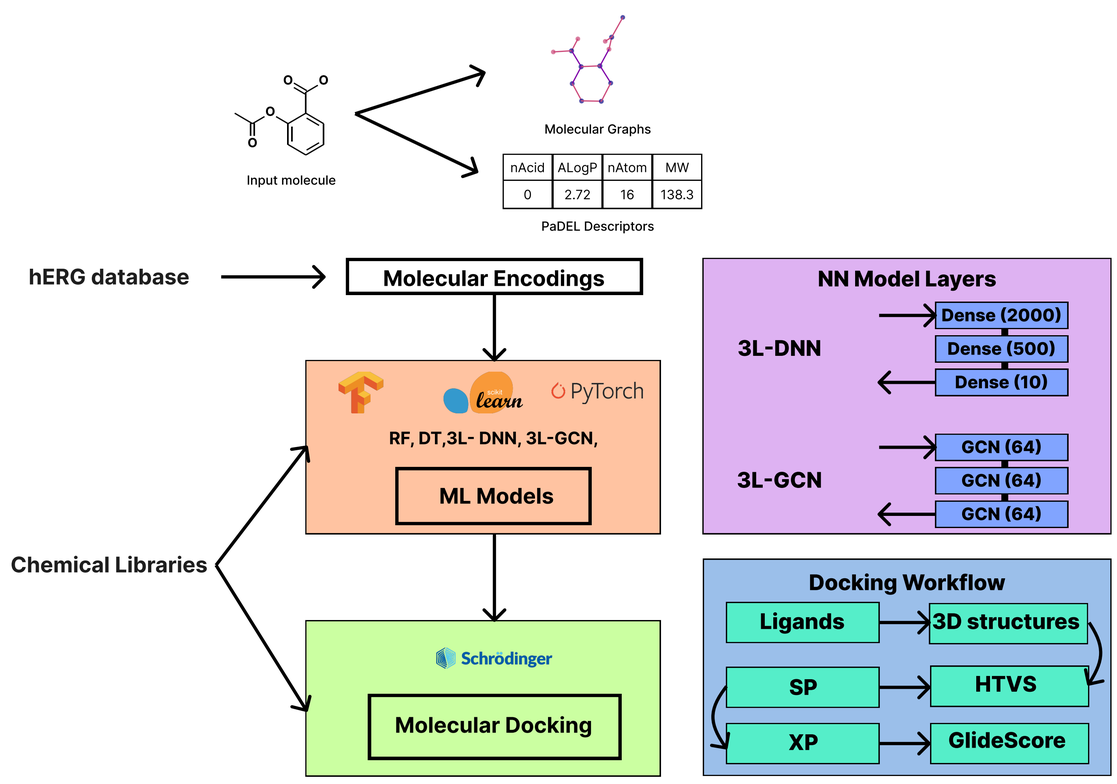


Figure 1. Workflow of the ML and Docking methods.

Results

ML Models

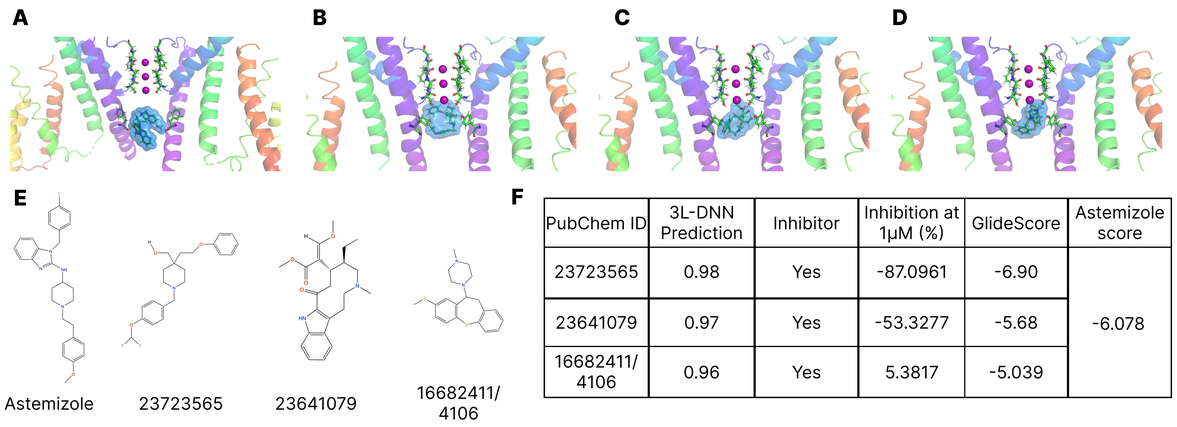
- Metrics used are standard classification metrics, obtained from the results of the confusion matrix.
- 3L-DNN best performing overall, due to high AUC-ROC and F1 score.

Model	Accuracy	AU ROC	F1	Precision	Recall	Specificity	TP	FP	TN	FN
DT	0.93	0.63	0.27	0.27	0.28	0.97	852	2360	69137	2156
RF	**0.96**	0.56	0.21	**0.75**	0.12	**0.99**	360	122	71375	2648
3L-DNN	0.87	**0.80**	**0.31**	0.20	**0.73**	0.88	2198	8813	62684	810
3L-GCN	0.85	0.66	0.20	0.12	0.45	0.87	1037	7265	49223	1225

Figure 2. Table of metrics of the classification models. Best performance indicated in bold.

Docking

- Top 3 predicted molecules from the 3L-DNN test set used for docking to hERG (RSCB: 7CN1).
- GlideScore of top predicted ligands close to that of Astemizole.
- However consensus docking needed to confirm results.



PubChem ID	3L-DNN Prediction	Inhibitor	Inhibition at 1μM (%)	GlideScore	Astemizole score
23723565	0.98	Yes	-87.0961	-6.90	-6.078
23641079	0.97	Yes	-53.3277	-5.68	
16682411/ 4106	0.96	Yes	5.3817	-5.039	

Figure 3. Results of docking of top 3 predicted molecules from 3L-DNN test set. (*A*) Docked Astemizole used as comparison. (*B*) Top pose of CID 23723565. (C) Top pose of CID 23641079. (D) Top pose of CID 16682411. (E) Structures of Astemizole and predicted ligands. (F) Top three predicted molecules with the highest likelihood of inhibition and resulting docking scores.

Conclusion

- 3L-DNN best model for prediction of hERGcentral dataset, able to distinguish between inhibitors and non-inhibitors.
- 3L-DNN can be tested on different datasets to improve metrics, and potentially predict hERG blockers.

References

1. Asai, T., Adachi, N., Moriya, T., Oki, H., Maru, T., Kawasaki, M., et al., Structure, (**2021**), 29 203–12 e4.
2. Du F, Yu H, Zou B, Babcock J, Long S, Li M., Assay Drug Dev Technol. (**2011**), 9(6):580-8. PMID: 22149888, PMCID: PMC3232635.